Expolortary Data Analysis

(EDA)

Asad Raza Virk

2024-09-18

Table of contents

1 Elements of Structured Data

- Data comes from many sources: sensor measurements, events, text, images, and videos.
- The Internet of Things (IoT) is spewing out streams of information. Much of this data is unstructured: images are a collection of pixels, with each pixel containing RGB (red, green, blue) color information.
- Texts are sequences of words and non-word characters, often organized by sections, subsections, and so on.
- Clickstreams are sequences of actions by a user interacting with an app or a web page.
- A major challenge of data science is to harness this torrent of raw data into actionable information.
- To apply the statistical concepts , unstructured raw data must be processed and manipulated into a structured form.

One of the commonest forms of structured data is a table with rows and columns—as data might emerge from a relational database or be collected for a study

1.1 Two basic types of Structured Data

1. Numeric
    - continuous
      such as wind speed or time duration
    - discrete
      such as the count of the occurrence of an event

2. Categorical (takes only fixed set of values)
    - Binary Binary data is an important special case of categorical data that takes on only one of two values, such as 0/1, yes/no, or true/false
    - Ordinal Ordinal data in which the categories are ordered; an example of this is a numerical rating (1, 2, 3, 4, or 5)

For the purposes of data analysis and predictive modeling, the data type is important to help determine the type of visual display, data analysis, or statistical model.

Data science software, such as R and Python, uses these data types to improve computational performance. More important, the data type for a variable determines how software will handle computations for that variable.

1.2 Key Terms for Data Types

1. Numeric Data that are expressed on a numeric scale.
    - Continuous Data that can take on any value in an interval. (Synonyms: Interval, float, numeric)
    - Discrete Data that can take on only integer values, such as counts. (Synonyms: integer, count)
2. Categorical Data that can take on only a specific set of values representing a set of possible categories. (Synonyms: enums, enumerated, factors, nominal)
    - Binary A special case of categorical data with just two categories of values, e.g., 0/1, true/false. (Synonyms: dichotomous, logical, indicator, boolean)
    - Ordinal Categorical data that has an explicit ordering. (Synonym: ordered factor)

- Key Ideas

- Data is typically classified in software by type.
- Data types include numeric (continuous, discrete) and categorical (binary, ordinal).
- Data typing in software acts as a signal to the software on how to process the data.

2 Rectangular and Non-Rectangualr Data

2.1 Rectangular Data

- The typical frame of reference for an analysis in data science is a rectangular data object, like a spreadsheet or database table.
- Rectangular data is the general term for a two-dimensional matrix with rows indicating records (cases) and columns indicating features (variables)
- Data frame is the specific format in R and Python.
- The data doesn't always start in this form: unstructured data (e.g., text) must be processed and manipulated so that it can be represented as a set of features in the rectangular data
- Data in relational databases must be extracted and put into a single table for most data analysis and modeling tasks.

2.1.1 Key Terms for Rectangular Data

1. Data frame Rectangular data (like a spreadsheet) is the basic data structure for statistical and machine learning models.
2. Feature A column within a table is commonly referred to as a feature. Synonyms: attribute, input, predictor, variable
3. Outcome Many data science projects involve predicting an outcome—often a yes/no outcome. The features are sometimes used to predict the outcome in an experiment or a study. Synonyms: dependent variable, response, target, output
4. Records A row within a table is commonly referred to as a record. Synonyms: case, example, instance, observation, pattern, sample

2.1.2 Data Frames and Indexes

Traditional database tables have one or more columns designated as an index, essentially a row number. This can vastly improve the efficiency of certain database queries.

In Python, with the pandas library, the basic rectangular data structure is a DataFrame object. By default, an automatic integer index is created for a DataFrame based on the order of the rows.

In pandas, it is also possible to set multilevel/hierarchical indexes to improve the efficiency of certain operations.

> Terminology Diferences Terminology for rectangular data can be confusing. Statisticians and data scientists use different terms for the same thing. For a statistician, predictor variables are used in a model to predict a response or dependent variable. For a data scientist, features are used to predict a target. One synonym is particularly confusing: computer scientists will use the term sample for a single row; a sample to a statistician means a collection of rows.

2.2 Non-Rectangular Data Structures

There are other data structures besides rectangular data.

- Time series data records successive measurements of the same variable. It is the raw material for statistical forecasting methods, and it is also a key component of the data produced by devices—the Internet of Things.
- Spatial data structures, which are used in mapping and location analytics, are more complex and varied than rectangular data structures. In the object representation, the focus of the data is an object (e.g., a house) and its spatial coordinates. The field view, by contrast, focuses on small units of space and the value of a relevant metric (pixel brightness, for example).
- Graph (or network) data structures are used to represent physical, social, and abstract relationships. For example, a graph of a social network, such as Facebook or LinkedIn, may represent connections between people on the network. Distribution hubs connected by roads are an example of a physical network. Graph structures are useful for certain types of problems, such as network optimization and recommender systems.

  > Graphs in Statistics In computer science and information technology, the term graph typically refers to a depiction of the connections among entities, and to the underlying data structure. In statistics, graph is used to refer to a variety of plots

and visualizations, not just of connections among entities, and the term applies only to the visualization, not to the data structure.

3 Estimates of Location of Data

- Variables with measured or count data might have thousands of distinct values.
- A basic step in exploring your data is getting a "typical value" for each feature (variable): an estimate of where most of the data is located (i.e., its central tendency).

3.1 Key Terms for Estimates of Location

- Mean
  - The sum of all values divided by the number of values.
    - Synonym
      - average
- Weighted mean
  - The sum of all values times a weight divided by the sum of the weights.
    - Synonym
      - weighted average
- Median
  - The value such that one-half of the data lies above and below.
    - Synonym
      - 50th percentile
- Percentile
  - The value such that P percent of the data lies below.
    - Synonym
      - quantile
- Weighted median
  - The value such that one-half of the sum of the weight
- Trimmed mean
  - The average of all values after dropping a fixed number of extreme values.

        * Synonym
            · truncated mean

- Robust
    - Not sensitive to extreme values.
        * Synonym
            · resistant

- Outlier
    - A data value that is very different from most of the data.
        * Synonym
            · extreme value

Metrics and Estimates

Statisticians often use the term estimate for a value calculated from the data at hand, to draw a distinction between what we see from the data and the theoretical true or exact state of affairs. Data scientists and business analysts are more likely to refer to such a value as a metric

3.2 Mean

3.2.1 Mean

- The mean is the sum of all values divided by the number of values.

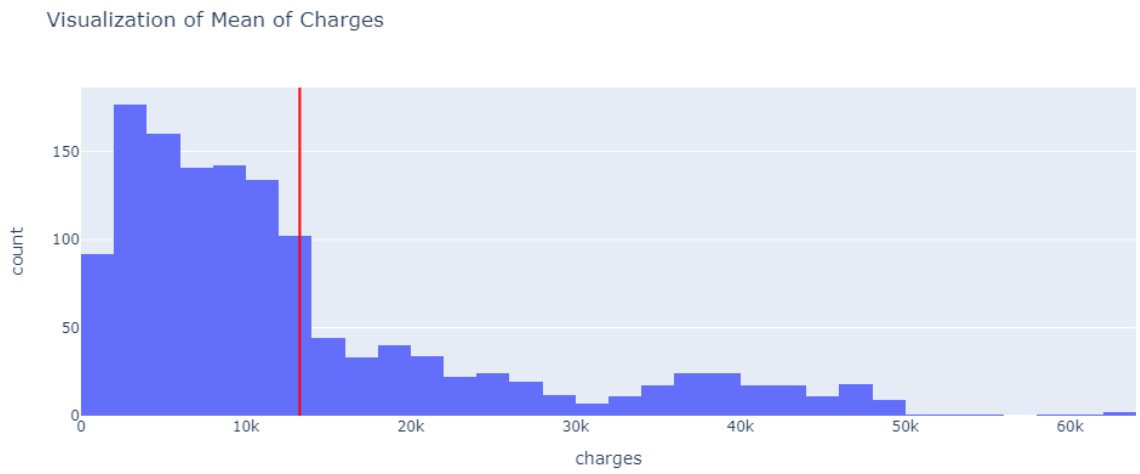


Figure 1: Mean

3.2.2 Trimmed Mean

- A variation of the mean which calculate the mean by dropping a fixed number of sorted values at each end and then taking an average of the remaining values
- A trimmed mean eliminates the influence of extreme values
- The trimmed mean is a robust measure of central tendency, as it is less affected by outliers than the mean
- The trimmed mean is calculated by first sorting the data in ascending order, then dropping a fixed number
- For example, in international diving the top score and bottom score from five judges are dropped, and the final score is the average of the scores from the three remaining judges. This makes it difficult for a single judge to manipulate the score, perhaps to favor their country's contestant.
- Trimmed means are widely used, and in many cases are preferable to using the ordinary mean

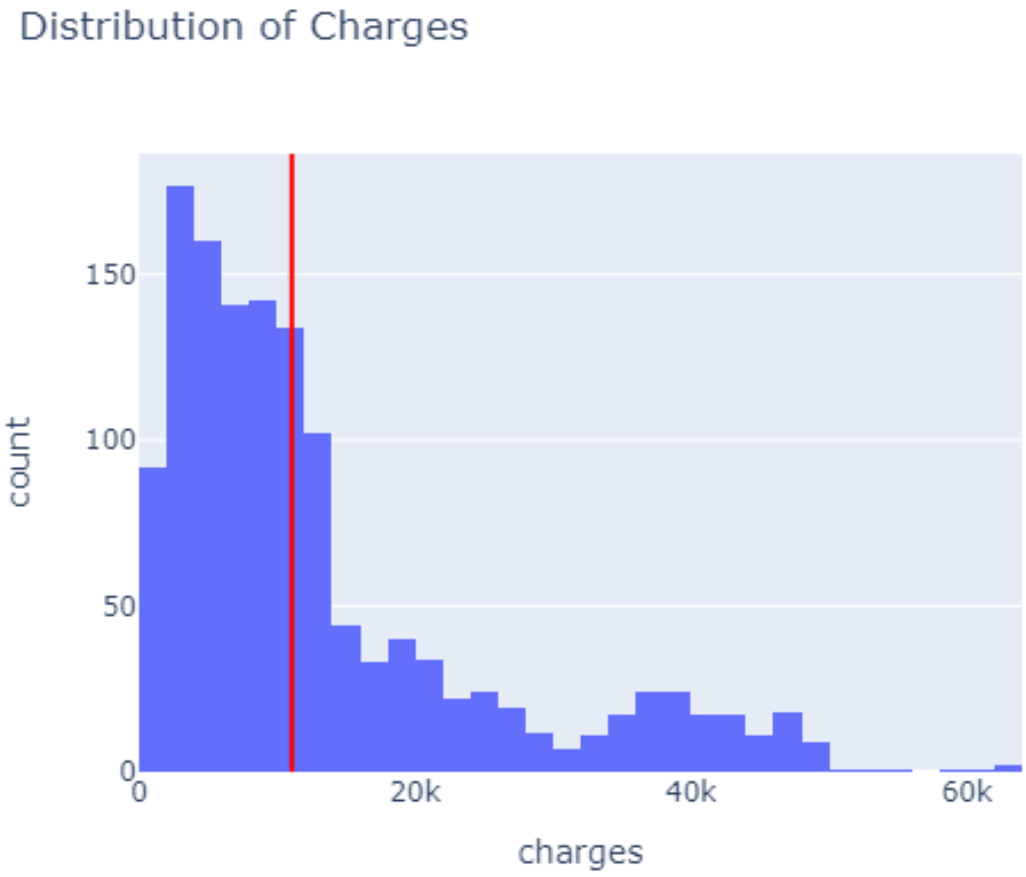
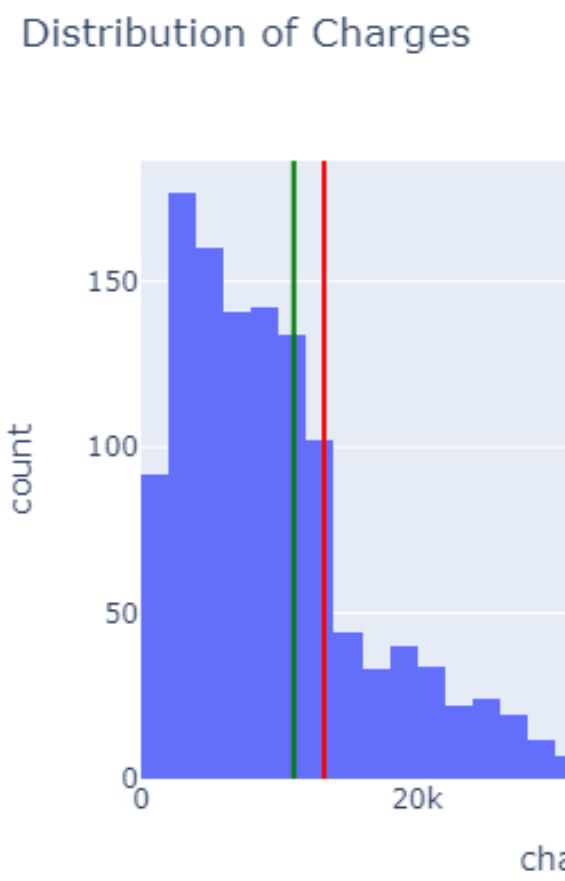


Figure 2: Trimmed Mean



Mean and Trimmed Mean (Mean = Red, Trimmed Mean = Greed)

3.3 Outliers

- outliers (extreme cases) could skew the results
- An outlier is any value that is very distant from the other values in a data set
- The exact definition of an outlier is somewhat subjective
- Outliers can be either high or low values
- Being an outlier in itself does not make a data value invalid or erroneous
- outliers are often the result of data errors such as mixing data of different units (kilometers versus meters) or bad readings from a sensor.
- When outliers are the result of bad data, the mean will result in a poor estimate of location, while the median will still be valid
- In any case, outliers should be identified and are usually worthy of further investigation.

Anomaly Detection

In contrast to typical data analysis, where outliers are sometimes informative and sometimes a nuisance, in anomaly detection the points of interest are the outliers, and the greater mass of data serves primarily to define the "normal" against which anomalies are measured.

The median is not the only robust estimate of location. In fact, a trimmed mean is widely used to avoid the influence of outliers. For example, trimming the bottom and top 10% (a common choice) of the data will provide protection against outliers in all but the smallest data sets. The trimmed mean can be thought of as a compromise between the median and the mean: it is robust to extreme values in the data, but uses more data to calculate the estimate for location.

Weighted mean is available with NumPy. For weighted median, we can use the specialized package wquantiles

3.3.1 Key Ideas

- The basic metric for location is the mean, but it can be sensitive to extreme values (outlier).
- Other metrics (median, trimmed mean) are less sensitive to outliers and unusual distributions and hence are more robust

4 Estimates of Variability

4.1 Variability

- Location is just one dimension in summarizing a feature
- A second dimension, variability, also referred to as dispersion, measures whether the data values are tightly clustered or spread out.
- At the heart of statistics lies variability:
  - measuring it
  - reducing it
  - distinguishing random from real variability
  - identifying the various sources of real variability
  - making decisions in the presence of it
- Variability is often measured using the range, interquartile range (IQR), variance, or standard deviation

4.1.1 Key Terms for Variability Metrics

- Deviations
  - The difference between the observed values and the estimate of location.
    - Synonyms
      - errors, residuals
- Variance
  - The sum of squared deviations from the mean divided by n – 1 where n is the number of data values.
    - Synonym
      - mean-squared-error (MSE)

Step-by-Step Example of Sample Variance Calculation:

1. Data Points: {3, 5, 7, 9, 11}

2. Calculate the Mean:
   Mean ($\bar{x}$) = (3 + 5 + 7 + 9 + 11) / 5 = 7

3. Subtract the Mean from Each Data Point (Deviations):
   (3 - 7) = -4, (5 - 7) = -2, (7 - 7) = 0, (9 - 7) = 2, (11 - 7) = 4
   Deviations: {-4, -2, 0, 2, 4}

4. Square Each Deviation:
   (-4)^2 = 16, (-2)^2 = 4, (0)^2 = 0, (2)^2 = 4, (4)^2 = 16
   Squared Deviations: {16, 4, 0, 4, 16}

5. Sum of Squared Deviations:
   16 + 4 + 0 + 4 + 16 = 40

6. Divide by n-1 (n=5):
   Variance = 40 / (5 - 1) = 40 / 4 = 10

Sample Variance = 10

Figure 3: Example

- Standard deviation
  - The square root of the variance.
    - Synonyms
      - root mean squared error (RMSE)
- Mean absolute deviation
  - The mean of the absolute values of the deviations from the mean.
    - Synonyms
      - l1-norm, Manhattan norm

Step-by-Step Example of Mean Absolute Deviation (MAD) Calculation:

Data points: {2, 4, 6, 8, 10}

1. Calculate the Mean:
   Mean ($\bar{x}$) = (2 + 4 + 6 + 8 + 10) / 5 = 6

2. Find the Absolute Deviations from the Mean:
   |2 - 6| = 4
   |4 - 6| = 2
   |6 - 6| = 0
   |8 - 6| = 2
   |10 - 6| = 4
   Absolute deviations: {4, 2, 0, 2, 4}

3. Calculate the Mean of the Absolute Deviations:
   MAD = (4 + 2 + 0 + 2 + 4) / 5 = 12 / 5 = 2.4

Mean Absolute Deviation (MAD) = 2.4

Figure 4: Example

- Median absolute deviation from the median
  - The median of the absolute values of the deviations from the median

Step-by-Step Example of Median Absolute Deviation (MAD) Calculation:

Data points: {2, 4, 6, 8, 10}

1. Find the Median of the Dataset:
   Median = 6 (the middle value in the sorted dataset)

2. Find the Absolute Deviations from the Median:
   |2 - 6| = 4
   |4 - 6| = 2
   |6 - 6| = 0
   |8 - 6| = 2
   |10 - 6| = 4
   Absolute deviations: {4, 2, 0, 2, 4}

3. Find the Median of the Absolute Deviations:
   Sorted absolute deviations: {0, 2, 2, 4, 4}
   Median of absolute deviations = 2

Median Absolute Deviation (MAD) = 2

Figure 5: Example

- Range
  - The difference between the largest and the smallest value in a data set.
- Order statistics
  - Metrics based on the data values sorted from smallest to biggest.
    - Synonym
      - ranks
- Percentile
  - The value such that P percent of the values take on this value or less and (100–P) percent take on this value or more.
    - Synonym
      - quantile

- Interquartile range
  - The difference between the 75th percentile and the 25th percentile.
    - Synonym
      - IQR

4.2 Standard Deviation and Related Estimates

- The most widely used estimates of variation are based on the differences, or deviations, between the estimate of location and the observed data.
- For a set of data {1, 4, 4}, the mean is 3 and the median is 4. The deviations from the mean are the differences: 1 – 3 = –2, 4 – 3 = 1, 4 – 3 = 1.
- These deviations tell us how dispersed the data is around the central value
- One way to measure variability is to estimate a typical value for these deviations.Averaging the deviations themselves would not tell us much—the negative deviations offset the positive ones. In fact, the sum of the deviations from the mean is precisely zero.
- Instead, a simple approach is to take the average of the absolute values of the deviations from the mean. In the preceding example, the absolute value of the deviations is {2 1 1}, and their average is (2 + 1 + 1) / 3 = 1.33. This is known as the mean absolute deviation
- The best-known estimates of variability are the variance and the standard deviation, which are based on squared deviations. The variance is an average of the squared deviations, and the standard deviation is the square root of the variance
- The standard deviation is much easier to interpret than the variance since it is on the same scale as the original data. Still, with its more complicated and less intuitive formula, it might seem peculiar that the standard deviation is preferred in statistics over the mean absolute deviation. It owes its preeminence to statistical theory: mathematically, working with squared values is much more convenient than absolute values, especially for statistical models.

  here is always some discussion of why we have n – 1 in the denominator in the variance formula, instead of n, leading into the concept of degreesof freedom. This distinction is not important since n is generally large enough that it won't make much difference whether you divide by n or n – 1. But in case you are interested, here is the story. It is based on the premise that you want to make estimates about a population, based

on a sample. If you use the intuitive denominator of n in the variance formula, you will underestimate the true value of the variance and the standard deviation in the population. This is referred to as a biased estimate. However, if you divide by n – 1 instead of n, the variance becomes an unbiased estimate. To fully explain why using n leads to a biased estimate involves the notion of degrees of freedom, which takes into account the number of constraints in computing an estimate. In this case, there are n – 1 degrees of freedom since there is one constraint: the standard deviation depends on calculating the sample mean. For most problems, data scientists do not need to worry about degrees of freedom.

- Neither the variance, the standard deviation, nor the mean absolute deviation is robust to outliers and extreme values
- The variance and standard deviation are especially sensitive to outliers since they are based on the squared deviations.
- A robust estimate of variability is the median absolute deviation from the median or MAD
- Like the median, the MAD is not influenced by extreme val- ues.
- It is also possible to compute a trimmed standard deviation analogous to the trimmed mean > The variance, the standard deviation, the mean absolute deviation, and the median absolute deviation from the median are not equivalent estimates, even in the case where the data comes from a normal distribution. In fact, the standard deviation is always greater than the mean absolute deviation, which itself is greater than the median absolute deviation. Sometimes, the median absolute deviation is multiplied by a constant scaling factor to put the MAD on the same scale as the standard deviation in the case of a normal distribution. The commonly used factor of 1.4826 means that 50% of the normal distribution fall within the range ±MAD

The purpose of calculating deviations in the values is to understand how each individual value differs from the average (mean) value. This helps in several ways:

- Measure of Variability: Deviations provide a measure of the spread or variability in the data. Large deviations indicate that the charges are spread out over a wide range, while small deviations suggest that the charges are clustered closely around the mean.
- Identify Outliers: By examining the deviations, you can identify outliers or unusual data points that are significantly different from the mean.

- Statistical Analysis: Deviations are a fundamental component in various statistical analyses, such as calculating the variance and standard deviation, which are key measures of data dispersion.
- Data Visualization: Plotting the deviations can help visualize the distribution of the data, making it easier to identify patterns, trends, and anomalies.

In statistical analysis, variance and deviation are closely related concepts that measure the spread or dispersion of a dataset. Here's how they are related:

- Deviation:
  - Deviation refers to the difference between each data point and the mean of the dataset.
  - Deviations can be positive or negative, depending on whether the data point is above or below the mean.
  - The sum of deviations from the mean is always zero, as positive deviations cancel out negative deviations
- Variance:
  - Variance is a measure of how much the data points in a dataset vary from the mean.
  - It is calculated as the average of the squared deviations from the mean.
  - Squaring the deviations ensures that all values are positive and gives more weight to larger deviations.
  - The variance is always non-negative and is zero if all data points are identical
- Relationship:
  - Variance is essentially the mean of the squared deviations.
  - While deviations provide a measure of individual differences from the mean, variance provides a single value that summarizes the overall dispersion of the dataset.
  - Variance is used to calculate the standard deviation, which is the square root of the variance and provides a measure of dispersion in the same units as the original data.

In summary, deviations are the building blocks for calculating variance, and variance provides a comprehensive measure of the spread of the data based on these deviations.

This code will output a table with the columns for value, mean, deviation, and variance, and it will also plot the deviations. Adjust the df[‘charges’] list with your actual data if needed.

Value	Mean	Deviation	Variance	Squared Deviation
16884.924	13270.422265141257	3614.5017348587426	146542766.49354792	13064622.79129686
1725.5523	13270.422265141257	-11544.869965141257	146542766.49354792	133284022.51202069
4449.462	13270.422265141257	-8820.960265141257	146542766.49354792	77809339.99920091
21984.47061	13270.422265141257	8714.048344858744	146542766.49354792	75934638.55653541
3866.8552	13270.422265141257	-9403.567065141257	146542766.49354792	88427073.54860935
3756.6216	13270.422265141257	-9513.800665141256	146542766.49354792	90512403.0960422
8240.5896	13270.422265141257	-5029.832665141257	146542766.49354792	25299216.639322
7281.5056	13270.422265141257	-5988.916665141256	146542766.49354792	35867122.822006665
6406.4107	13270.422265141257	-6864.011565141256	146542766.49354792	47114654.76639292
28923.13692	13270.422265141257	15652.714654858744	146542766.49354792	245007476.0664297

Figure 6: Example

The variance is a single value that summarizes the overall dispersion of the dataset. It is not specific to individual data points but rather describes the dataset as a whole. Therefore, when you include the variance in the DataFrame, it is the same for every row because it represents the same overall measure of variability for the entire dataset.

If you want to include the variance in the DataFrame, it should be shown as a single value, not repeated for each data point. However, if you want to show the squared deviations (which contribute to the variance), you can include those instead.

In statistical analysis, the standard deviation and variance are both measures of the spread or dispersion of a dataset. They are closely related but differ in how they express this dispersion.

- Variance
- Definition: Variance measures the average squared deviations from the mean.

- Units: The units of variance are the square of the units of the original data. For example, if the data is in meters, the variance will be in square meters.
- Formula: Variance = Σ(xi -)^2 / (n - 1), where
  - xi is each data point,
  - is the mean of the dataset,
  - n is the number of data points, and
  - Σ denotes the sum of the squared deviations.
- Standard Deviation
- Definition: Standard deviation is the square root of the variance. It provides a measure of dispersion in the same units as the original data.
- Units: The units of standard deviation are the same as the units of the original data. For example, if the data is in meters, the standard deviation will also be in meters.
- Relationship Mathematical Relationship: The standard deviation is the square root of the variance.
- Interpretation:
  - Variance gives a measure of how data points spread out from the mean, but because it uses squared units, it can be less intuitive.
  - Standard deviation, being in the same units as the data, is often more interpretable and is commonly used to describe the spread of the data.
- Example If you have a dataset, the variance tells you the average of the squared differences from the mean, while the standard deviation tells you how much the data points typically deviate from the mean in the original units of the data.

4.3 Estimates Based on Percentiles

A different approach to estimating dispersion is based on looking at the spread of the sorted data.Statistics based on sorted (ranked) data are referred to as order statistics.

- The most basic measure is the range: the difference between the largest and smallest numbers. The minimum and maximum values themselves are useful to know and are helpful in identifying outliers, but the range is extremely sensitive to outliers and not very useful as a general measure of dispersion in the data.
- To avoid the sensitivity to outliers, we can look at the range of the data after dropping values from each end Formally, these types of estimates are based on differences between percentiles
- In a data set, the Pth percentile is a value such that at least P percent of the values take on this value or less and at least (100 – P) percent of the values take on this value or more. For example, to find the 80th percentile, sort the data. Then, starting with the smallest value, proceed 80 percent of the way to the largest value. Note that the median is the same thing as the 50th percentile. The percentile is essentially the same as a quantile, with quantiles indexed by fractions (so the .8 quantile is the same as the 80th percentile).
- A common measurement of variability is the difference between the 25th percentile and the 75th percentile, called the interquartile range (or IQR).
- Software can have slightly differing approaches that yield different answers
- For very large data sets, calculating exact percentiles can be computationally very expensive since it requires sorting all the data values. Machine learning and statistical software use special algorithms, such as [Zhang-Wang-2007], to get an approximate percentile that can be calculated very quickly and is guaranteed to have a certain accuracy.

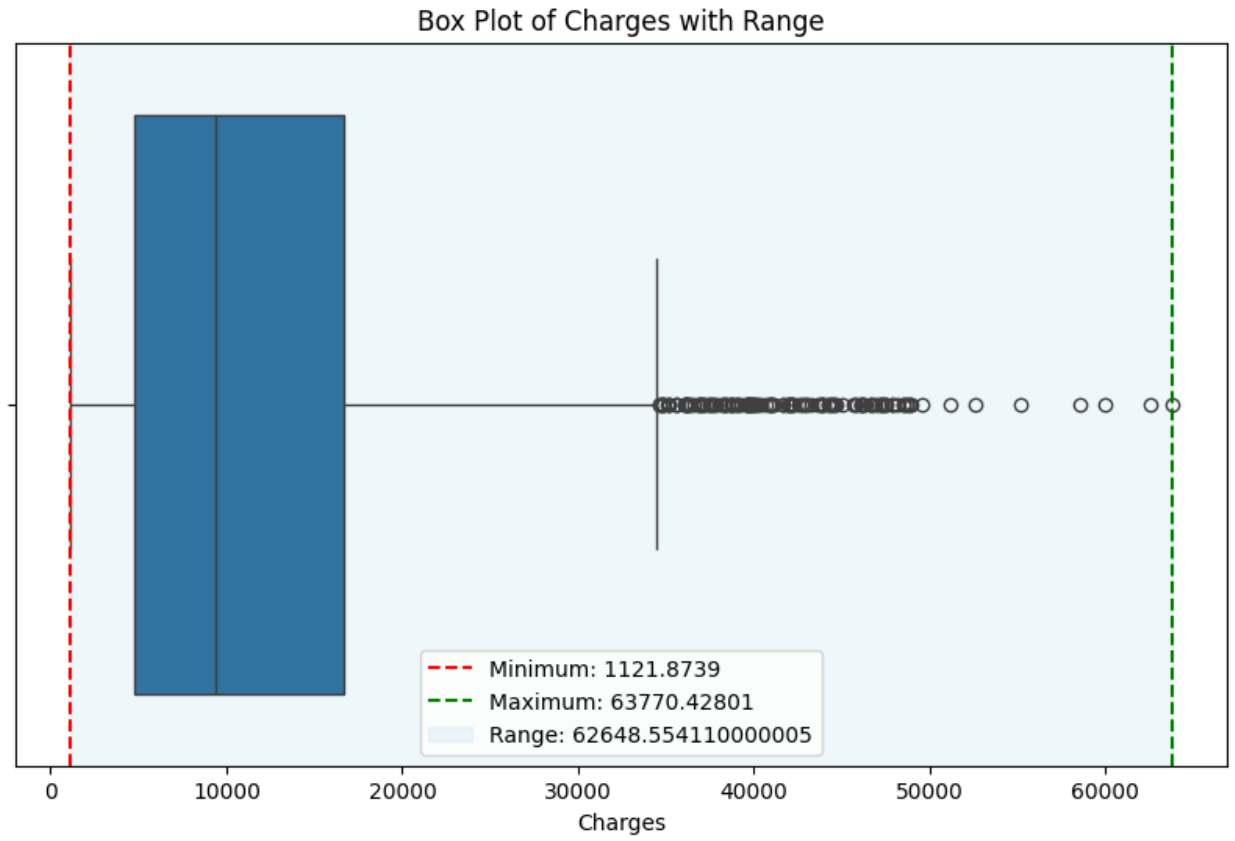


Figure 7: Range

Order statistics are metrics based on the data values sorted from smallest to largest. They provide insights into the distribution and spread of the data. Here are some common order statistics:

- Minimum: The smallest value in the dataset.
- Maximum: The largest value in the dataset.
- Median: The middle value when the data is sorted. If the dataset has an even number of observations, the median is the average of the two middle values.
- Quartiles: Values that divide the dataset into four equal parts.
  - First Quartile (Q1): The median of the lower half of the dataset (25th percentile).
  - Second Quartile (Q2): The median of the dataset (50th percentile).
  - Third Quartile (Q3): The median of the upper half of the dataset (75th percentile).
- Percentiles: Values that divide the dataset into 100 equal parts. For example, the 90th percentile is the value below which 90% of the data falls.

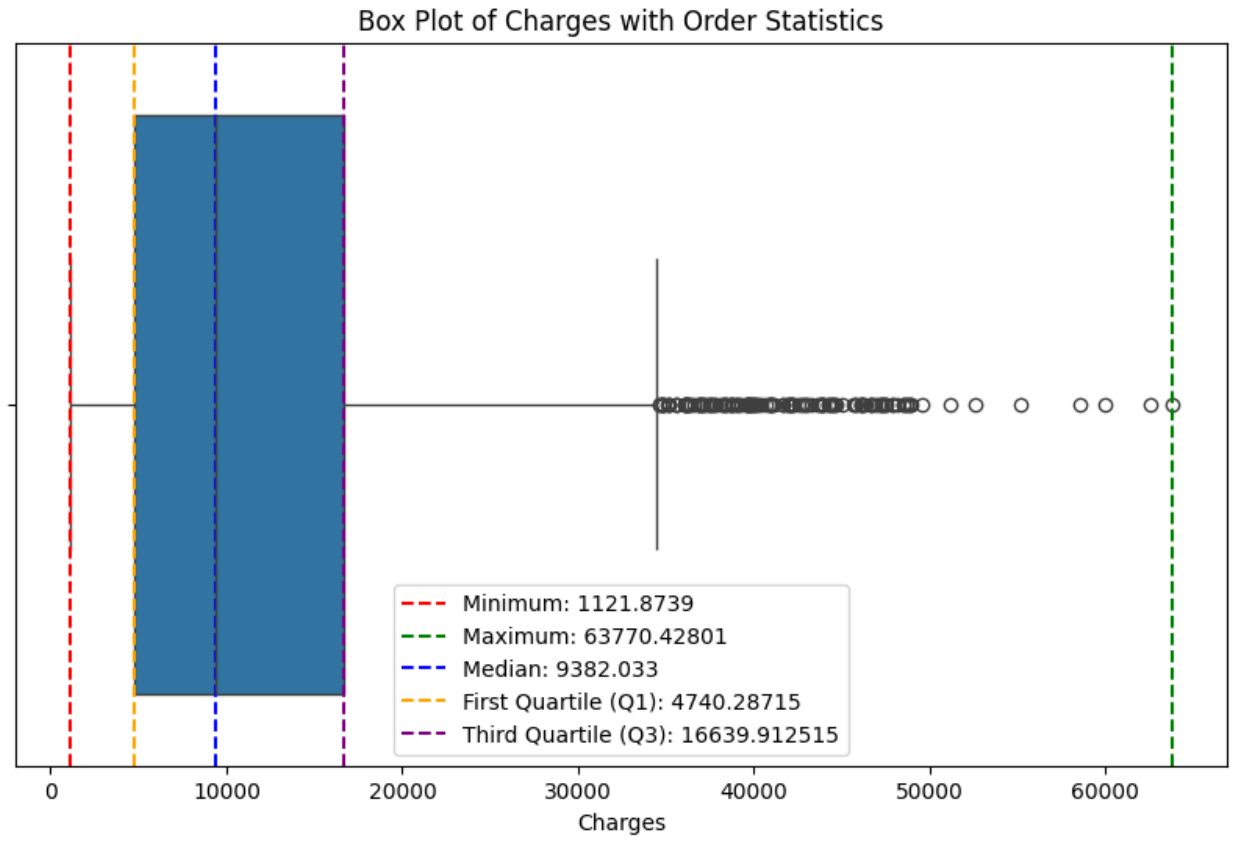


Figure 8: Percentile

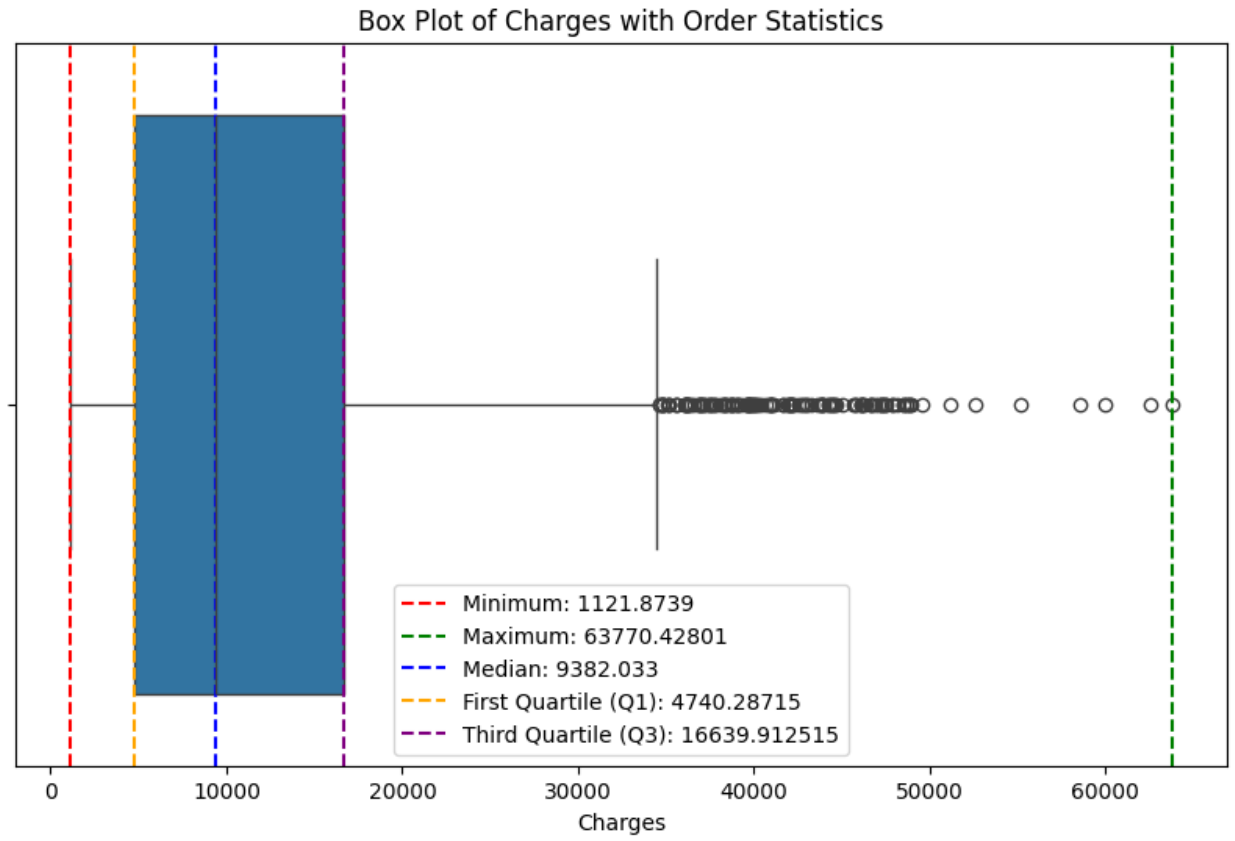


Figure 9: IQR

4.3.1 Key Ideas

- Variance and standard deviation are the most widespread and routinely reported statistics of variability.
- Both are sensitive to outliers.
- More robust metrics include mean absolute deviation, median absolute deviation from the median, and percentiles (quantiles).

5 Exploring the Data Distribution

Each of the estimates sums up the data in a single number to describe the location or variability of the data. It is also useful to explore how the data is distributed overall.

5.1 Key Terms for Exploring the Distribution

- Boxplot
  - A plot introduced by Tukey as a quick way to visualize the distribution of data.
    - Synonym
      - box and whiskers plot
- Frequency table
  - A tally of the count of numeric data values that fall into a set of intervals (bins).
- Histogram
  - A plot of the frequency table with the bins on the x-axis and the count (or pro- portion) on the y-axis. While visually similar, bar charts should not be confused with histograms.
- Density plot
  - A smoothed version of the histogram, often based on a kernel density estimate.

5.2 Percentiles and Boxplots

5.2.1 Percentiles

Percentiles can be used to measure the spread of the data. Percentiles are also valuable for summarizing the entire distribution. It is common to report the quartiles (25th, 50th, and 75th per-centiles) and the deciles (the 10th, 20th, ..., 90th percentiles). Percentiles are especially valuable for summarizing the tails (the outer range) of the distribution. Popular culture has coined the term one-percenters to refer to the people in the top 99th percentile of wealth.

	Quantile	Value
0	0.005	1145.11
1	0.1	2346.53
2	0.15	3171.84
3	0.25	4740.29
4	0.5	9382.03
5	0.75	16639.9
6	0.85	24990.2
7	0.9	34831.7
8	0.95	41181.8

Figure 10: Percentile

5.2.2 Boxplots

Boxplots introduced by Tukey [Tukey-1977], are based on percentiles and give a quick way to visualize the distribution of data.

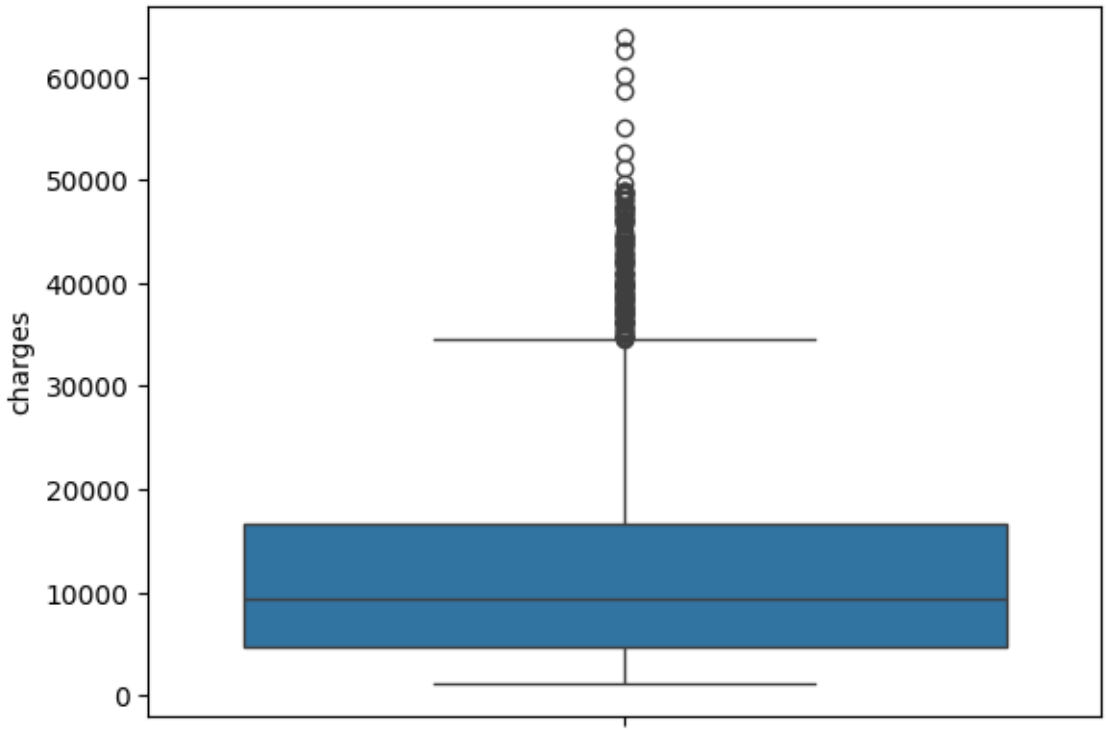


Figure 11: Boxplot

- The top and bottom of the box are the 75th and 25th percentiles, respectively
- The median is shown by the horizontal line in the box
- The whiskers are the two lines outside the box that extend to the highest and lowest observations that are within 1.5 * IQR from the upper and lower quartiles
- Any points outside this range are plotted as individual points
- There are many variations of a boxplot
- By default, the R function extends the whiskers to the furthest point beyond the box, except that it will not go beyond 1.5 times the IQR. Matplotlib uses the same implementation; other software may use a different rule.
- Any data outside of the whiskers is plotted as single points or circles (often considered outliers).
- Pandas provides a number of basic exploratory plots for data frame; one of them is boxplots

5.3 Frequency Tables and Histograms

5.3.1 Frequency Tables

- A frequency table of a variable divides up the variable range into equally spaced segments and tells us how many values fall within each segment.
- The function pandas.cut creates a series that maps the values into the segments. Using the method value_counts, we get the frequency table

	Charges	Count
0	(0, 10000]	712
1	(10000, 20000]	353
2	(20000, 30000]	111
3	(30000, 40000]	83
4	(40000, 50000]	72

Figure 12: Frequency Table

It is important to include the empty bins; the fact that there are no values in those bins is useful information. It can also be useful to experiment with different bin sizes. If they are too large, important features of the distribution can be obscured.If they are too small, the result is too granular, and the ability to see the bigger picture is lost.

Both frequency tables and percentiles summarize the data by creat-ing bins. In general, quartiles and deciles will have the same count in each bin (equal-count bins), but the bin sizes will be different. The frequency table, by contrast, will have different counts in the bins (equal-size bins), and the bin sizes will be the same.

5.3.2 Histogram

A histogram is a way to visualize a frequency table, with bins on the x-axis and the data count on the y-axis. Pandas supports histograms for data frames with the DataFrame.plot.hist method. Use the keyword argument bins to define the number of bins. The various plot meth- ods return an axis object that allows further fine-tuning of the visualization using Matplotlib:

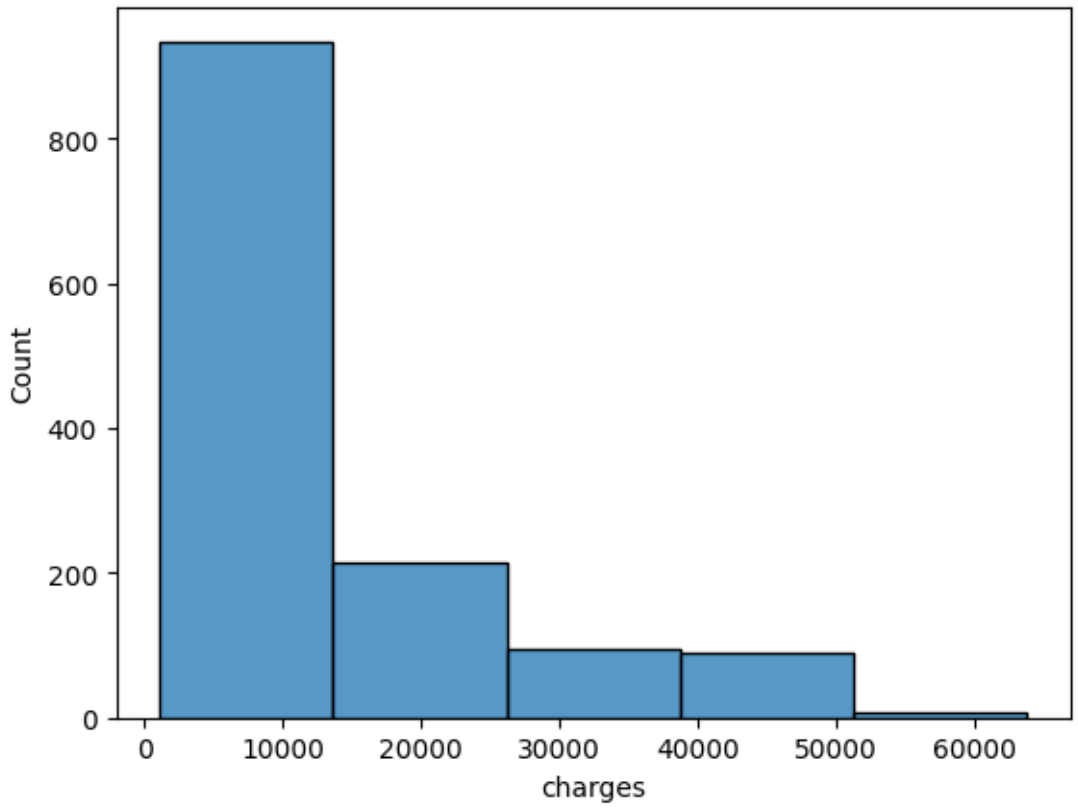


Figure 13: Histogram

Histograms are plotted such that:

- Empty bins are included in the graph.
- Bins are of equal width.
- The number of bins (or, equivalently, bin size) is up to the user.
- Bars are contiguous—no empty space shows between bars, unless there is an empty bin.

Statistical Moments

In statistical theory, location and variability are referred to as the first and second moments of a distribution. The third and fourth moments are called skewness and kurtosis. Skewness refers to whether the data is skewed to larger or smaller values, and kurtosis indicates the propensity of the data to have extreme values. Generally, metrics are not used to measure skewness and kurtosis; instead, these are discovered through visual displays

5.4 Density Plots and Estimates

- Related to the histogram is a density plot, which shows the distribution of data values as a continuous line.
- A density plot can be thought of as a smoothed histogram, although it is typically computed directly from the data through a kernel density estimate
- pandas provides the density method to create a density plot. Use the argument bw_method to control the smoothness of the density curve

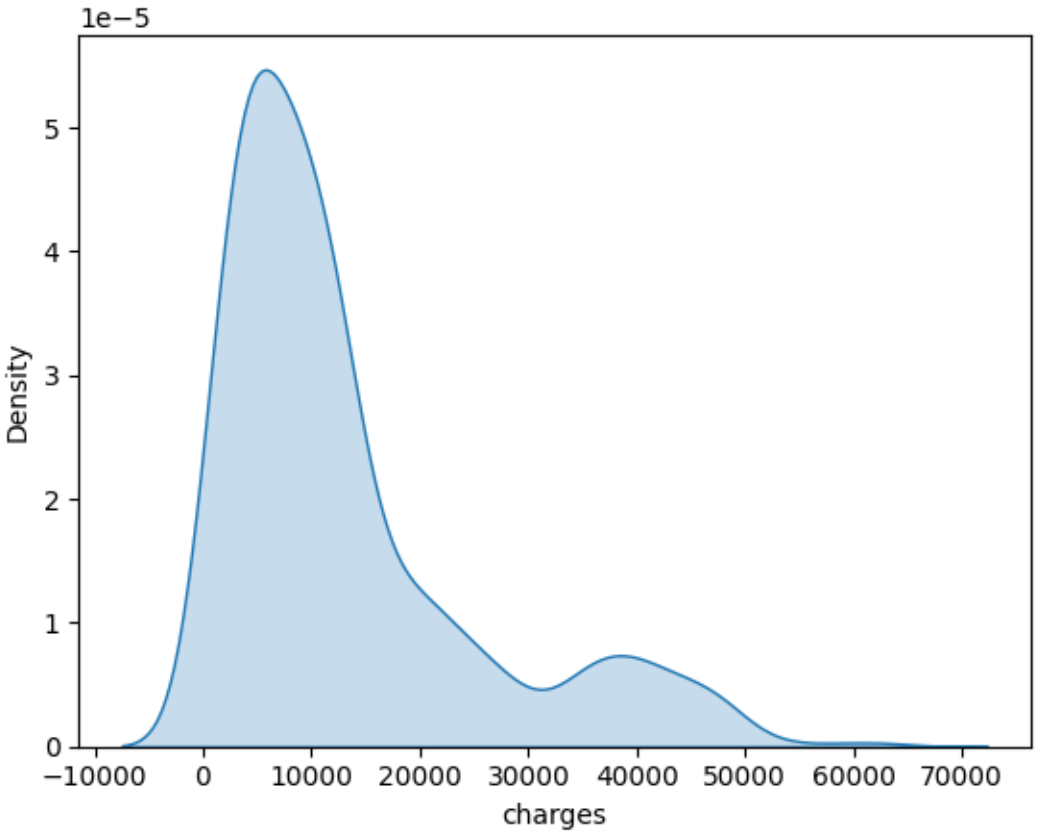


Figure 14: Density Plot

A key distinction from the histogram is the scale of the y-axis: a density plot corresponds to plotting the histogram as a proportion rather than counts. Note that the total area under the density curve = 1, and instead of counts in bins you calculate areas under the curve between any two points on the x-axis, which correspond to the proportion of the distribution lying between those two points

Density Estimation

Density estimation is a rich topic with a long history in statistical literature. The density estimation methods in pandas and scikit-learn also offer good implementations. For many data science problems, there is no need to worry about the various types of density estimates; it suffices to use the base functions.

5.5 Key Ideas

- A frequency histogram plots frequency counts on the y-axis and variable values on the x-axis; it gives a sense of the distribution of the data at a glance.
- A frequency table is a tabular version of the frequency counts found in a histogram.
- A boxplot—with the top and bottom of the box at the 75th and 25th percentiles, respectively—also gives a quick sense of the distribution of the data; it is often used in side-by-side displays to compare distributions.
- A density plot is a smoothed version of a histogram; it requires a function to esti- mate a plot based on the data (multiple estimates are possible, of course).

6 Exploring Binary and Categorical Data

For categorical data, simple proportions or percentages tell the story of the data. For binary data, the proportion of 1s (or 0s) is often the most important metric. For both types of data, the mode (the most common value) is often the most informative summary statistic.

6.1 Key Terms for Exploring Categorical Data

- Mode
  - The most commonly occurring category or value in a data set.
- Frequency
  - The number of times a value occurs in a data set.
- Proportion
  - The fraction of the data that takes on a particular value.
- Bar chart
  - A plot of the frequency or proportion for each category of a categorical variable.
- Pie chart
  - A plot that shows the proportion of cases that fall into each category of a categorical variable.
- Expected value
  - When the categories can be associated with a numeric value, this gives an average value based on a category's probability of occurrence
  - Mean of a probability distribution

Getting a summary of a binary variable or a categorical variable with a few categories is a fairly easy matter: we just figure out the proportion of 1s, or the proportions of the important categories.

Sex	Total	Smokers	%
female	662	115	17.372
male	676	159	23.521

Figure 15: Percentage of Smokers

6.2 Bar Charts

- A bar chart is a common way to represent categorical data. The height of each bar represents the frequency (or proportion) of each category.
- A common visual tool for displaying a single categorical variable
- Categories are listed on the x-axis, and frequencies or pro- portions on the y-axis

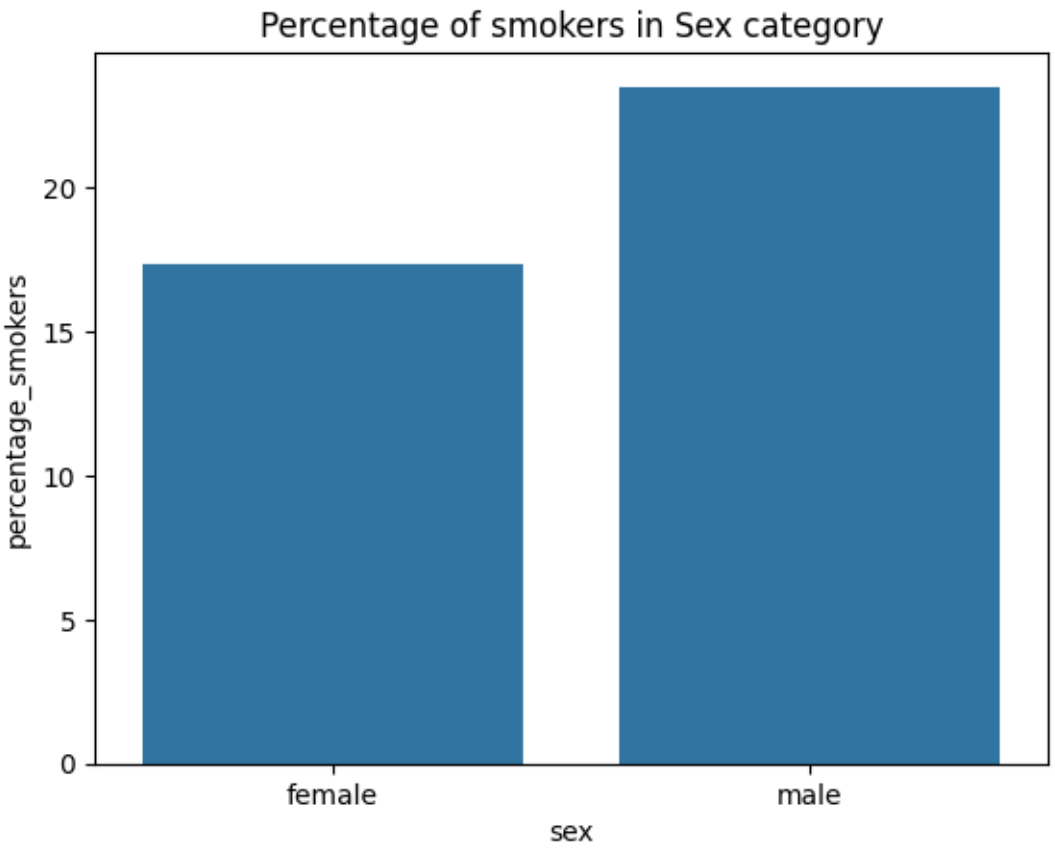


Figure 16: Bar Chart

- The bar chart is a simple and effective way to visualize categorical data. It is especially useful when the number of categories is small, and the data is not too granular. For larger numbers of categories, the chart can become cluttered and difficult to interpret.

Note that a bar chart resembles a histogram; in a bar chart the x-axis represents dif- ferent categories of a factor variable, while in a histogram the x-axis represents values of a single variable on a numeric scale. In a histogram, the bars are typically shown touching each other, with gaps indicating values that did not occur in the data. In a bar chart, the bars are shown separate from one another.

6.3 Pie Charts

- A pie chart is another way to visualize the distribution of a categorical variable. The size of each slice of the pie represents the proportion of each category.
- Pie charts are often used to show the relative sizes of the categories in a categorical variable

- The pie chart is a common way to visualize the distribution of a single categorical variable. It is especially useful when the number of categories is small and the data is not too granular. For larger numbers of categories, the chart can become cluttered and difficult to interpret.
- Pie charts are often criticized for being difficult to interpret, especially when there are many categories or when the categories are not ordered by size. In these cases, a bar chart is often a better choice.

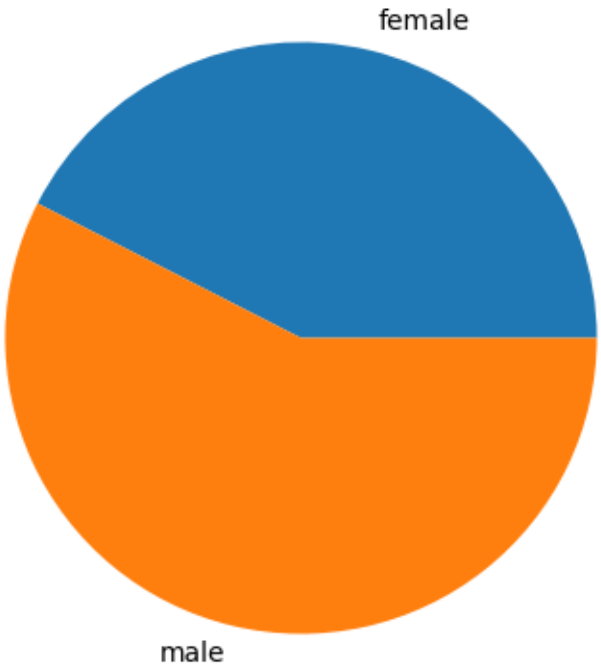


Figure 17: Pie Chart

Pie charts are an alternative to bar charts, although statisticians and data visualization experts generally eschew pie charts as less visually informative.

Numerical Data as Categorical Data

The frequency tables are based on binning the data. This implicitly converts the numeric data to an ordered factor. In this sense, histograms and bar charts are similar, except that the categories on the x-axis in the bar chart are not ordered. Converting numeric data to categorical data is an important and widely used step in data analysis since it reduces the complexity (and size) of the

data. This aids in the discovery of relationships between features, particularly at the initial stages of an analysis.

6.4 Mode

- The mode is the value—or values in case of a tie—that appears most often in the data.
- The mode is a simple summary statistic for categorical data, and it is generally not used for numeric data.
- The mode is especially useful for understanding the central tendency of categorical data, and it is often used in conjunction with bar charts and pie charts.

6.5 Expected Value

- A special type of categorical data is data in which the categories represent or can be mapped to discrete values on the same scale.
- In this case, the expected value is the average value of the variable, weighted by the probability of each category.
- The expected value is really a form of weighted mean: it adds the ideas of future expectations and probability weights, often based on subjective judgment. Expected value is a fundamental concept in business valuation and capital budgeting.

6.6 Probability

- The probability of a value occurring. Most people have an intuitive understanding of probability. For example, the probability of a fair coin landing heads is 0.5.
- In data science, probability is used to model uncertainty and randomness in data. It is a key concept in statistical inference, machine learning, and decision-making under uncertainty.
- Probability is often used to estimate the likelihood of an event occurring, given certain conditions or assumptions. It is expressed as a value between 0 and 1, where 0 indicates impossibility and 1 indicates certainty.

- In data analysis, probability is used to model the likelihood of different outcomes, such as the probability of a customer making a purchase, the probability of a machine failing, or the probability of a patient having a particular disease.
- Probability theory provides a mathematical framework for analyzing random events and making predictions based on data. It is a fundamental concept in data science and is used in various statistical models, such as Bayesian inference, logistic regression, and decision trees.
- Probability is also used to calculate expected values, which represent the average outcome of a random variable based on its probability distribution. Expected values are used in decision-making, risk assessment, and optimization problems.
- In summary, probability is a key concept in data science that is used to model uncertainty, make predictions, and analyze random events. It provides a foundation for statistical inference, machine learning, and decision-making under uncertainty.

6.7 Key Ideas

- For binary and categorical data, the mode is often the most informative summary statistic.
- Bar charts and pie charts are common ways to visualize categorical data.
- Expected value is a useful concept when the categories can be mapped to discrete values.
- Categorical data is typically summed up in proportions and can be visualized in a bar chart.
- Categories might represent distinct things (apples and oranges, male and female), levels of a factor variable (low, medium, and high), or numeric data that has been binned.
- Expected value is the sum of values times their probability of occurrence, often used to sum up factor variable levels.

7 Correlation

- Correlation is a measure of the strength and direction of the relationship between two variables. It is a key concept in statistics and data analysis, as it helps to identify patterns, trends, and associations in the data.
- Correlation is often used to determine whether two variables are related and to what extent. It is commonly used in predictive modeling, hypothesis testing, and feature selection.
- Exploratory data analysis in many modeling projects (whether in data science or in research) involves examining correlation among predictors, and between predictors and a target variable. Variables X and Y (each with measured data) are said to be posi- tively correlated if high values of X go with high values of Y, and low values of X go with low values of Y. If high values of X go with low values of Y, and vice versa, the variables are negatively correlated.
- Correlation is a statistical measure that ranges from -1 to 1. A correlation of 1 indicates a perfect positive relationship, a correlation of -1 indicates a perfect negative relationship, and a correlation of 0 indicates no relationship between the variables.

	age	bmi	children	charges
age	1.0	0.1092718815485351	0.04246899855884958	0.299008193330648
bmi	0.1092718815485351	1.0	0.012758900820673994	0.19834096883362912
children	0.04246899855884958	0.012758900820673994	1.0	0.06799822684790495
charges	0.299008193330648	0.19834096883362912	0.06799822684790495	1.0

Figure 18: Correlation

- The correlation coefficient is a measure of the strength and direction of the relationship between two variables. It ranges from -1 to 1, where:
  - 1 indicates a perfect positive relationship,
  - -1 indicates a perfect negative relationship, and
  - 0 indicates no relationship between the variables.

7.1 Key Terms for Correlation

- Correlation coefficient A metric that measures the extent to which numeric variables are associated with one another (ranges from –1 to +1).
- Correlation matrix A table where the variables are shown on both rows and columns, and the cell values are the correlations between the variables.
- Scatterplot A plot in which the x-axis is the value of one variable, and the y-axis the value of another.
- Covariance A measure of how changes in one variable are associated with changes in a second variable.

7.2 Correlation Coefficient

- More useful is a standardized variant
- Gives an estimate of the correlation between two variables that always lies on the same scale
- To compute Pearson's correlation coefficient, we multiply deviations from the mean for variable 1 times those for variable 2, and divide by the product of the standard deviations for the two variables
- The correlation coefficient always lies between +1 (perfect positive correlation) and –1 (perfect negative correlation); 0 indicates no correlation.
- Variables can have an association that is not linear, in which case the correlation coefficient may not be a useful metric.
- A table of correlation coefficients for all pairs of variables is called a correlation matrix.
- The correlation matrix is a square matrix that shows the correlation coefficients between all pairs of variables in a dataset. It is a useful tool for identifying relationships between variables and understanding the structure of the data.
- The correlation matrix is often used in exploratory data analysis to identify patterns, trends, and associations in the data. It can help to identify which variables are related and to what extent, which is useful for feature selection, predictive modeling, and hypothesis testing.

- The correlation matrix is a key tool in data analysis and is commonly used in statistics, machine learning, and data science to understand the relationships between variables in a dataset.

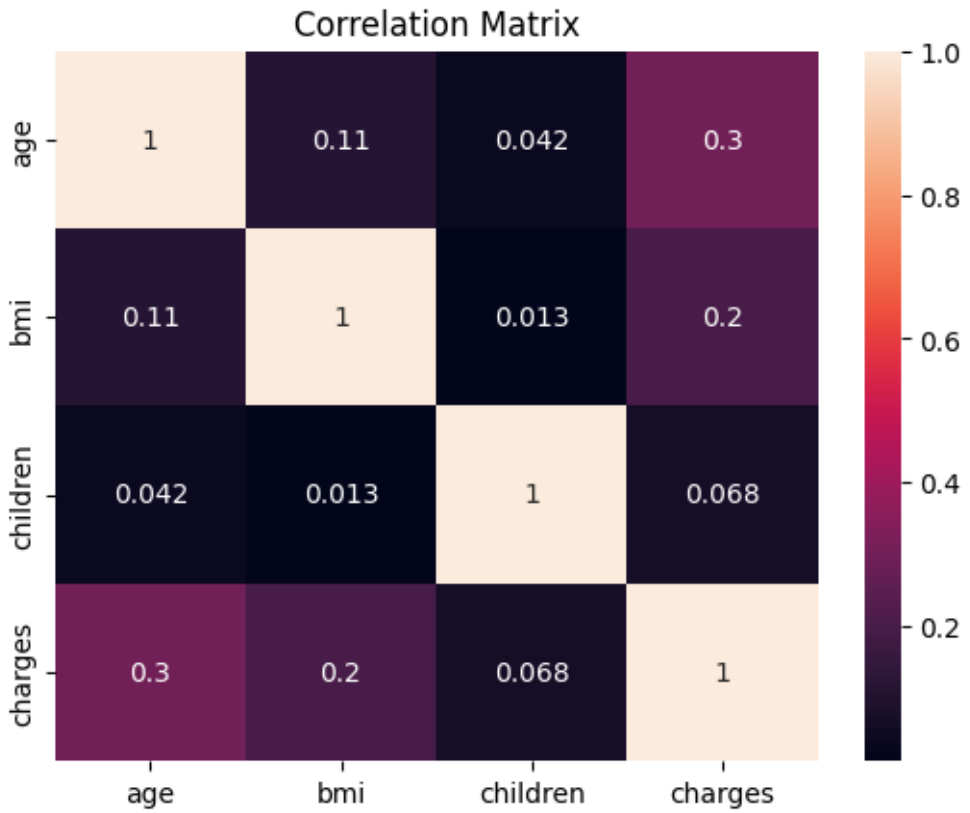


Figure 19: Correlation Matrix HeatMap

- Like the mean and standard deviation, the correlation coefficient is sensitive to outliers in the data
- Software packages offer robust alternatives to the classical correlation coefficient that are less sensitive to outliers.
- The methods in the scikit-learn module sklearn.covariance implement a variety of approaches

Other Correlation Estimates

Statisticians long ago proposed other types of correlation coefficients, such as Spearman's rho or Kendall's tau. These are correlation coefficients based on the rank of the data. Since they work with ranks rather than values, these estimates are robust to outliers and can handle certain types of nonlinearities. However, data scientists can generally stick to Pearson's correlation coefficient,

and its robust alternatives, for exploratory analysis. The appeal of rankbased estimates is mostly for smaller data sets and specific hypothesis tests.

7.3 Scatterplot

- The standard way to visualize the relationship between two measured data variables is with a scatterplot. The x-axis represents one variable and the y-axis another, and each point on the graph is a record.
- A scatterplot is a visual representation of the relationship between two variables. It is commonly used in data analysis to identify patterns, trends, and associations in the data.
- Scatterplots are useful for visualizing the relationship between two continuous variables. They help to identify correlations, outliers, and nonlinear relationships in the data.
- Scatterplots are often used in exploratory data analysis to understand the structure of the data and to identify potential relationships between variables.

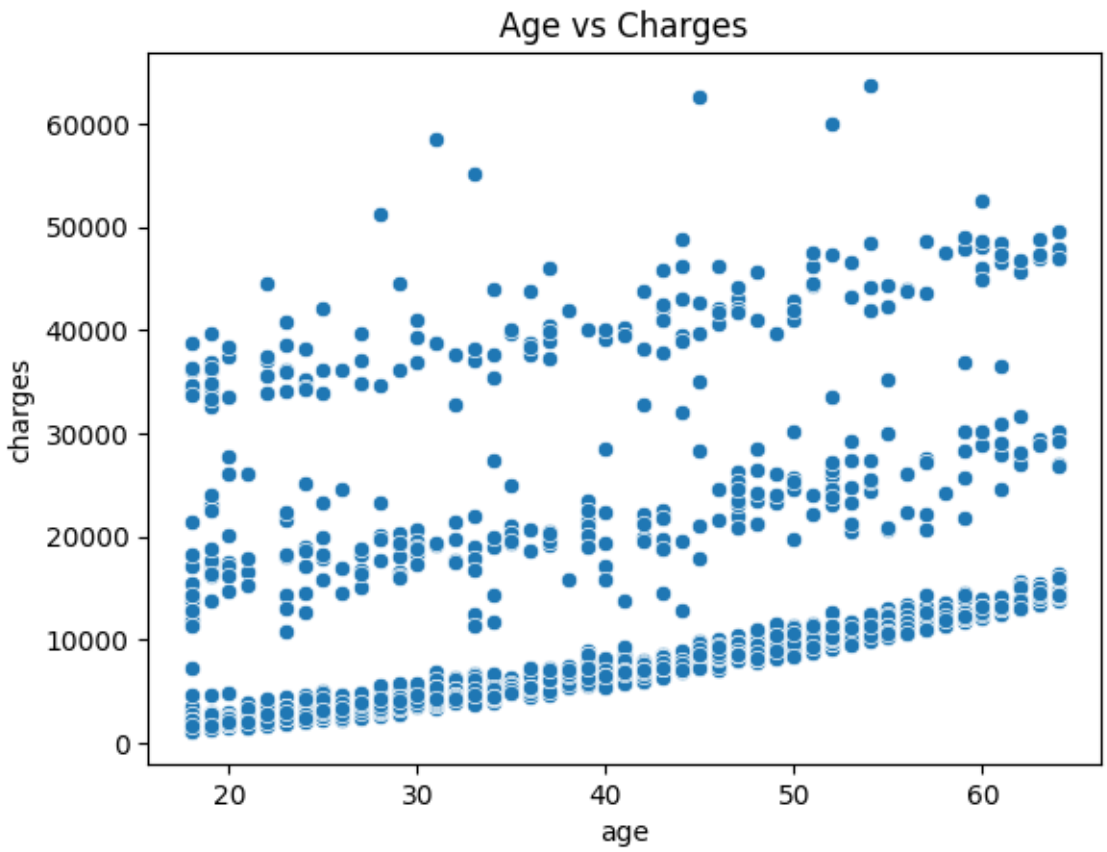


Figure 20: Scatterplot

7.4 Key Ideas

- The correlation coefficient measures the extent to which two paired variables (e.g., height and weight for individuals) are associated with one another.
- When high values of v1 go with high values of v2, v1 and v2 are positively associated.
- When high values of v1 go with low values of v2, v1 and v2 are negatively associated.
- The correlation coefficient is a standardized metric, so that it always ranges from –1 (perfect negative correlation) to +1 (perfect positive correlation).
- A correlation coefficient of zero indicates no correlation, but be aware that random arrangements of data will produce both positive and negative values for the correlation coefficient just by chance.

8 Exploring Two or More Variables

- Estimators like mean and variance look at variables one at a time (univariate analysis)
- Correlation analysis is an important method that compares two variables (bivariate analysis)
- In practice, we often need to look at more than two variables at a time (multivariate analysis)
- Multivariate analysis is a key part of data analysis and is used to identify patterns, trends, and relationships between multiple variables in a dataset.
- Multivariate analysis is used in various fields, such as statistics, machine learning, and data science, to understand complex relationships between variables and to make predictions based on multiple factors.

8.1 Key Terms for Exploring Two or More Variables

- Contingency table A tally of counts between two or more categorical variables.
- Hexagonal binning A plot of two numeric variables with the records binned into hexagons.
- Contour plot A plot showing the density of two numeric variables like a topographical map.
- Violin plot Similar to a boxplot but showing the density estimate.

Like univariate analysis, bivariate analysis involves both computing summary statistics and producing visual displays. The appropriate type of bivariate or multivariate analysis depends on the nature of the data: numeric versus categorical.

8.2 Hexagonal Binning and Contours(Plotting Numeric Versus Numeric Data)

- When you have a large number of data points, scatterplots can become too dense to interpret. One solution is to bin the data and plot the bins. A common approach is hexagonal binning, where the plot is divided into hexagons, and the number of points in each hexagon is counted.
- Hexagonal binning is a useful technique for visualizing the relationship between two numeric variables when the data is dense and a scatterplot is difficult to interpret.

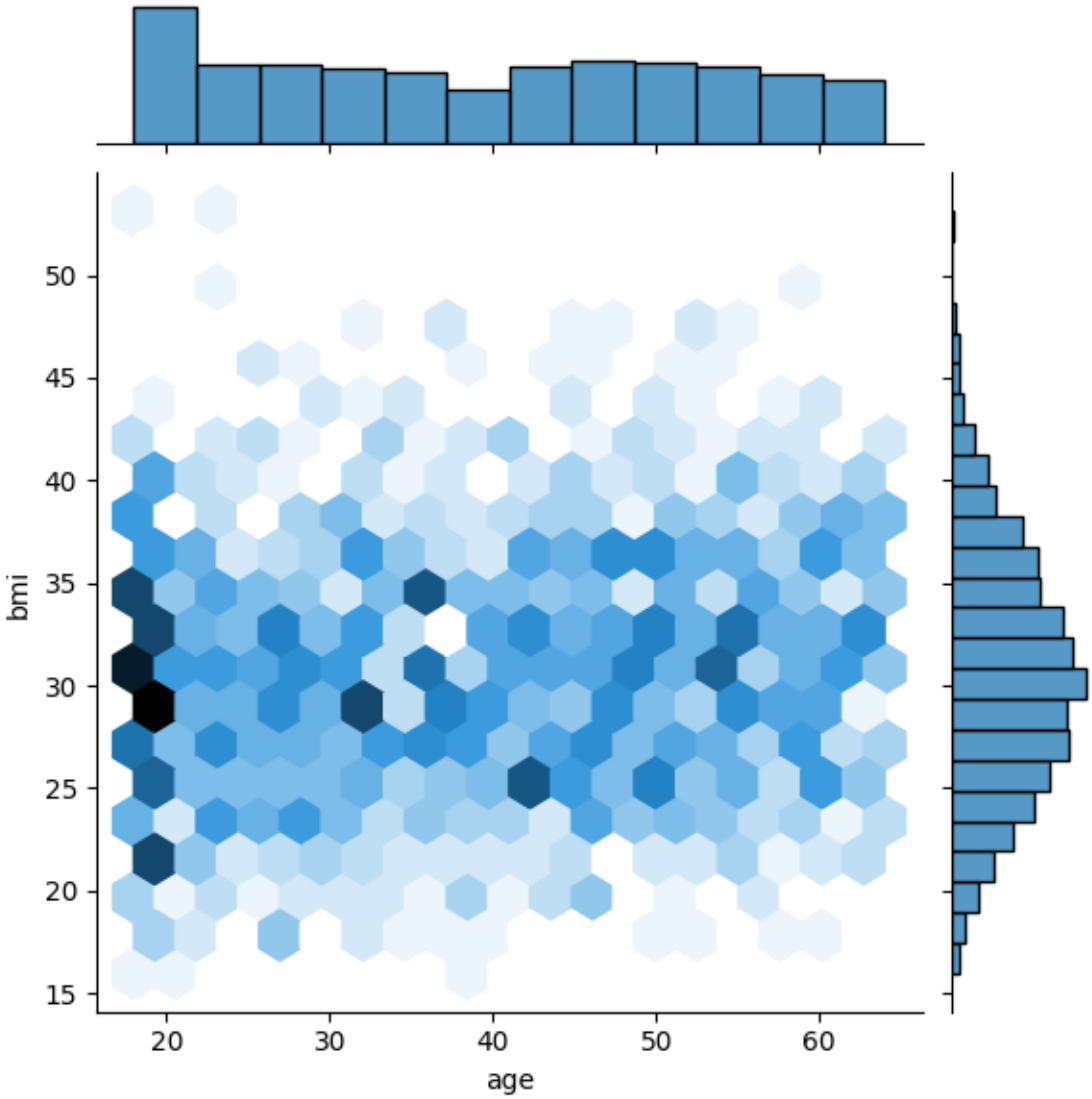


Figure 21: Hexagonal Binning

- Another approach is to use contour plots, which are similar to topographical maps. The density of the data is shown by the contours, with darker areas indicating higher density.
- Contour plots are useful for visualizing the relationship between two numeric variables and identifying patterns in the data.

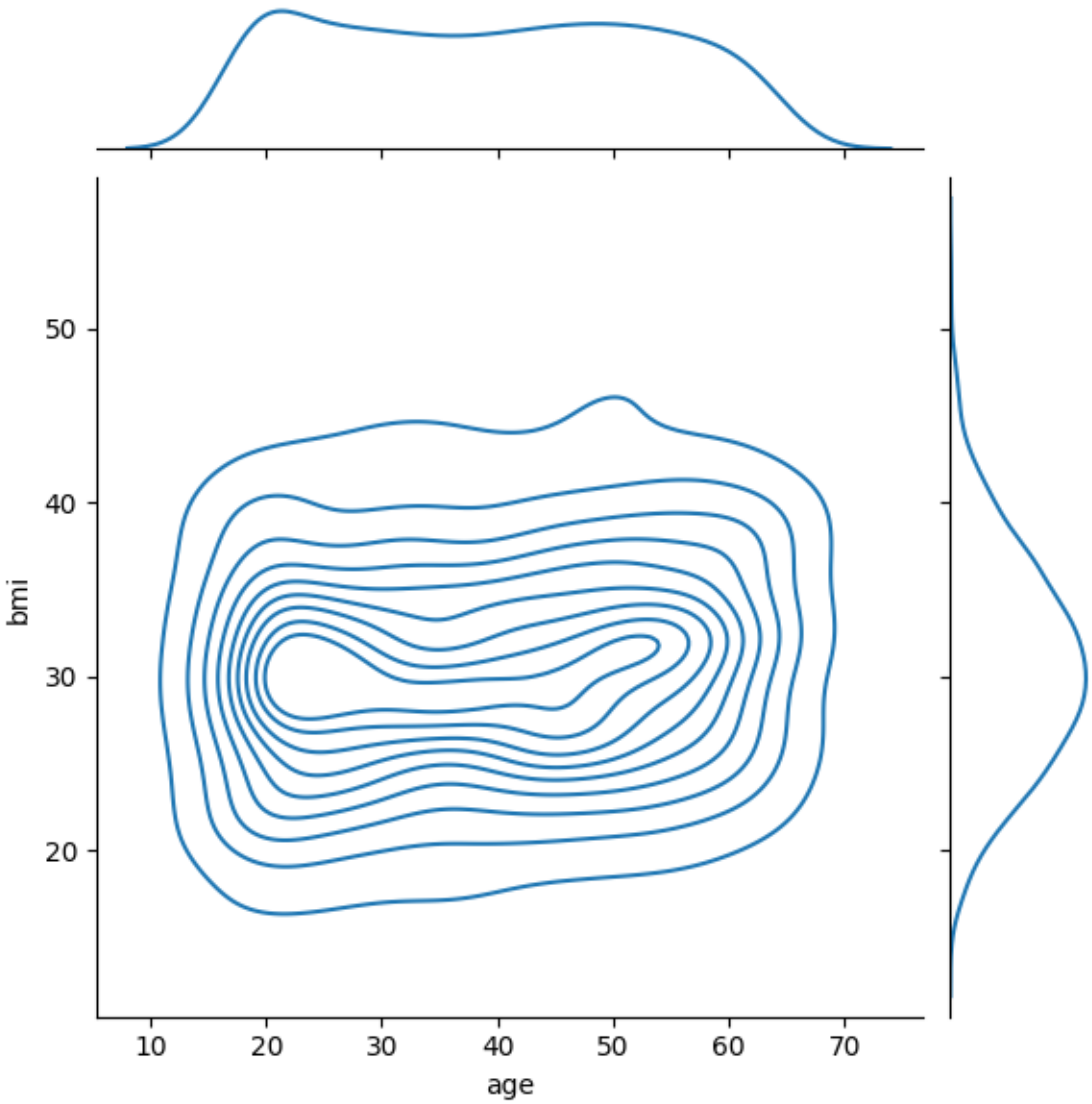


Figure 22: Contour Plot

The contours are essentially a topographical map to two variables; each contour band represents a specific density of points, increasing as one nears a “peak.”

Other types of charts are used to show the relationship between two numeric variables, including heat maps. Heat maps, hexagonal binning, and contour plots all give a visual representation of a two-dimensional density. In this way, they are natural analogs to histograms and density plots.

8.3 Two Categorical Variables

- For two categorical variables, a contingency table is a useful way to summarize the data. The table shows the counts of the data points that fall into each combination of categories.

	15.96	16.815	17.195	17.29	17.385	17.4	17.48	17.67	17.765	17.8
18	1	0	0	1	0	0	0	0	0	0
19	0	0	0	0	0	0	1	0	0	1
20	0	0	0	0	0	0	0	0	0	0
21	0	1	0	0	0	1	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	1	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0
26	0	0	1	0	0	0	0	1	0	0
27	0	0	0	0	0	0	0	0	0	0

Figure 23: Contingency Table

Contingency tables can look only at counts, or they can also include column and total percentages. Pivot tables in Excel are perhaps the most common tool used to create contingency tables. The pandas library in Python also has a pivot_table method that can be used to create contingency

tables.

8.4 Categorical and Numeric Data

- When one variable is categorical and the other is numeric, a boxplot is a useful way to visualize the data. The boxplot shows the distribution of the numeric variable for each category of the categorical variable.

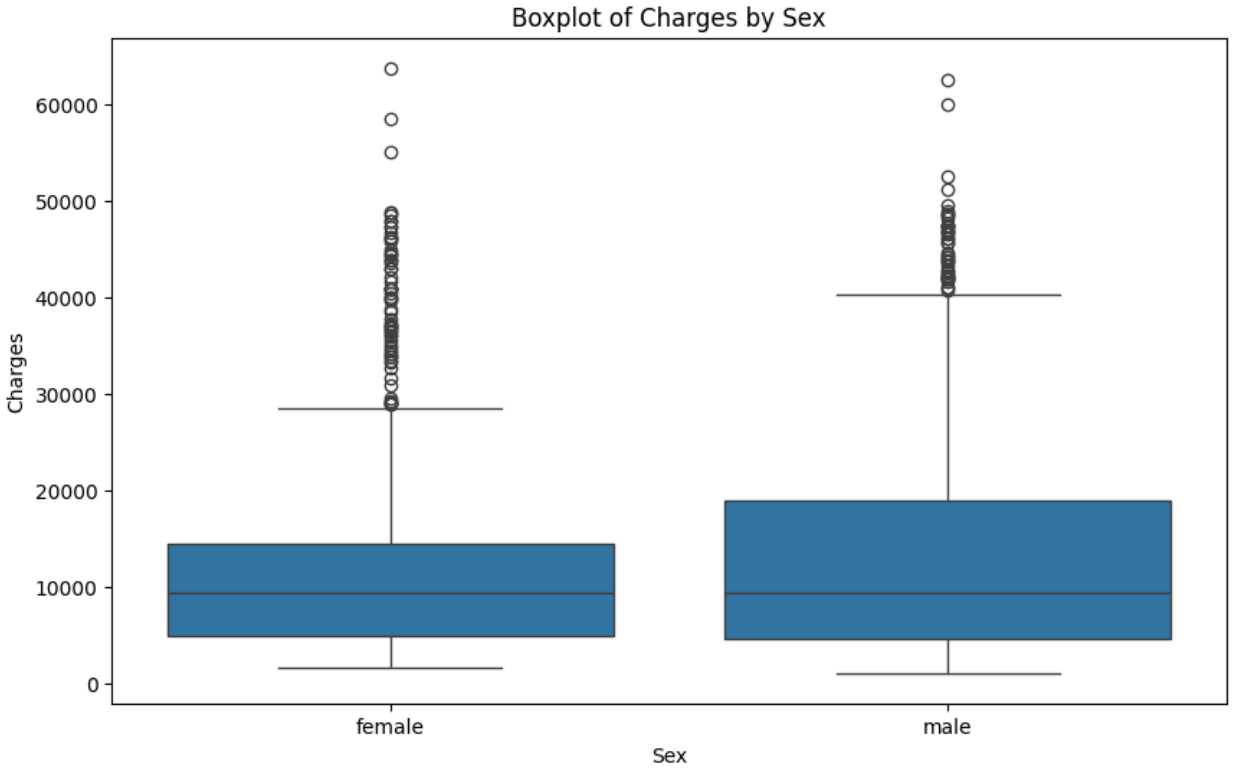


Figure 24: Boxplot

A violin plot is an enhancement to the boxplot and plots the density estimate with the density on the y-axis. The density is mirrored and flipped over, and the resulting shape is filled in, creating an image resembling a violin. The advantage of a violin plot is that it can show nuances in the distribution that aren't perceptible in a boxplot. On the other hand, the boxplot more clearly shows the outliers in the data.

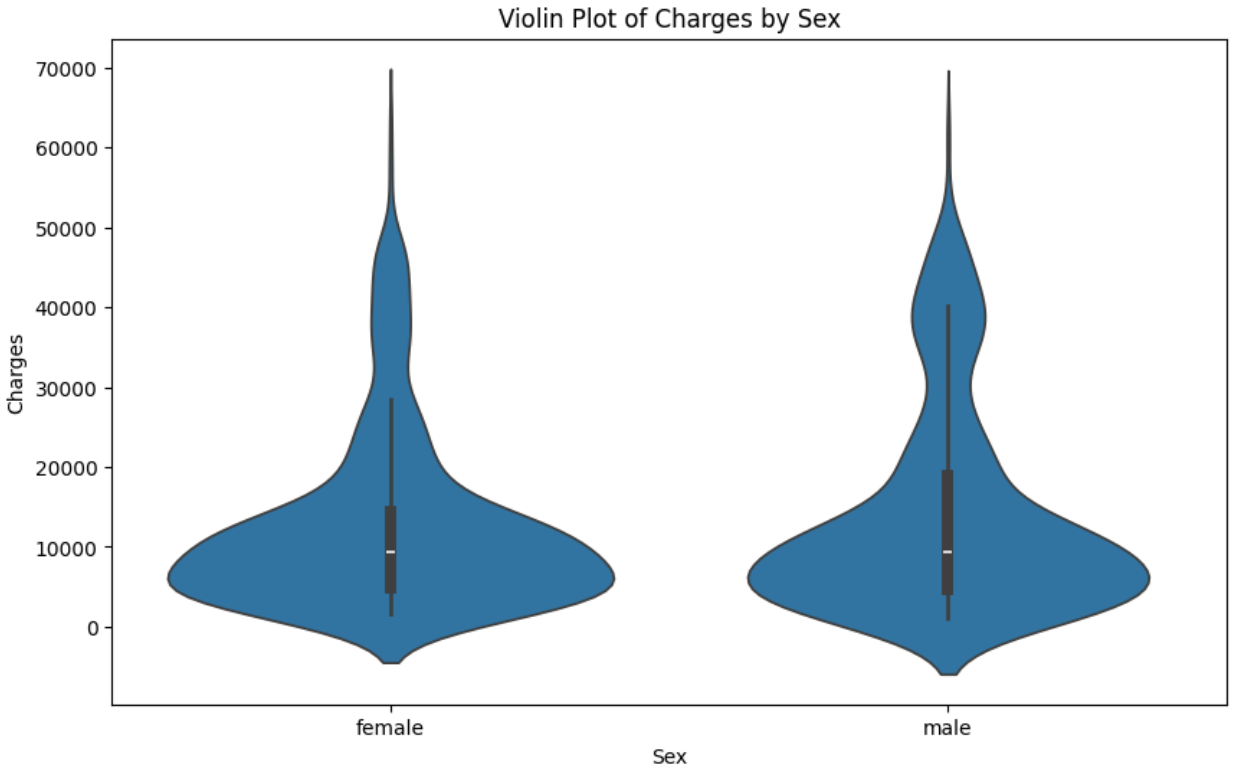


Figure 25: Violin Plot

8.5 Visualizing Multiple Variables

- When you have more than two variables, it is often useful to visualize the relationships between multiple variables simultaneously. One common approach is to use a pair plot, which shows scatterplots of all pairs of variables in a dataset.
- A pair plot is a grid of scatterplots showing the relationships between all pairs of variables in a dataset. It is a useful tool for visualizing the relationships between multiple variables and identifying patterns in the data.
- The types of charts used to compare two variables—scatterplots, hexagonal binning, and boxplots—are readily extended to more variables through the notion of conditioning

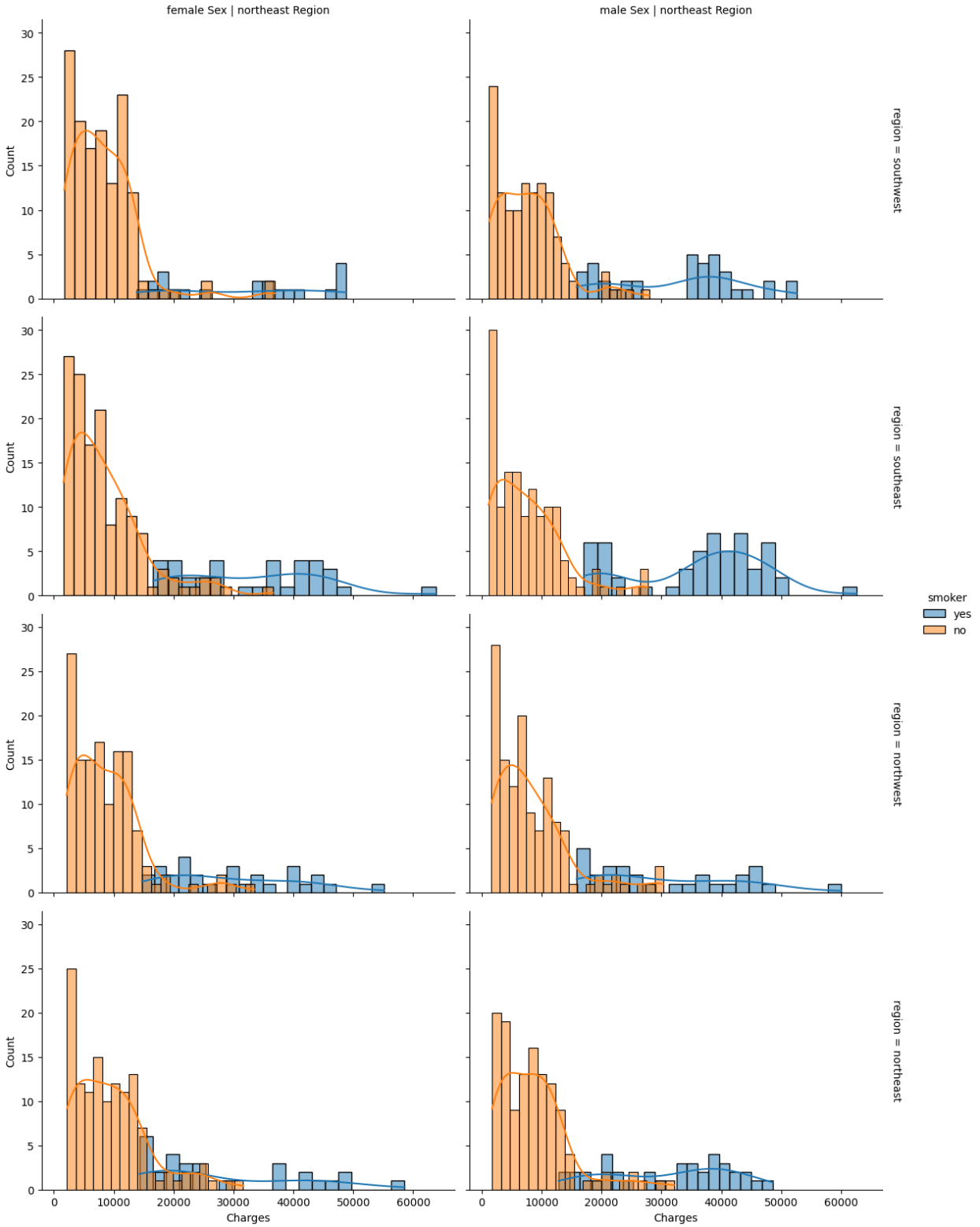


Figure 26: faceting

- Faceting is a technique that involves creating multiple plots, each showing a subset of the

data based on a categorical variable. It is a useful way to compare the relationships between multiple variables across different categories.

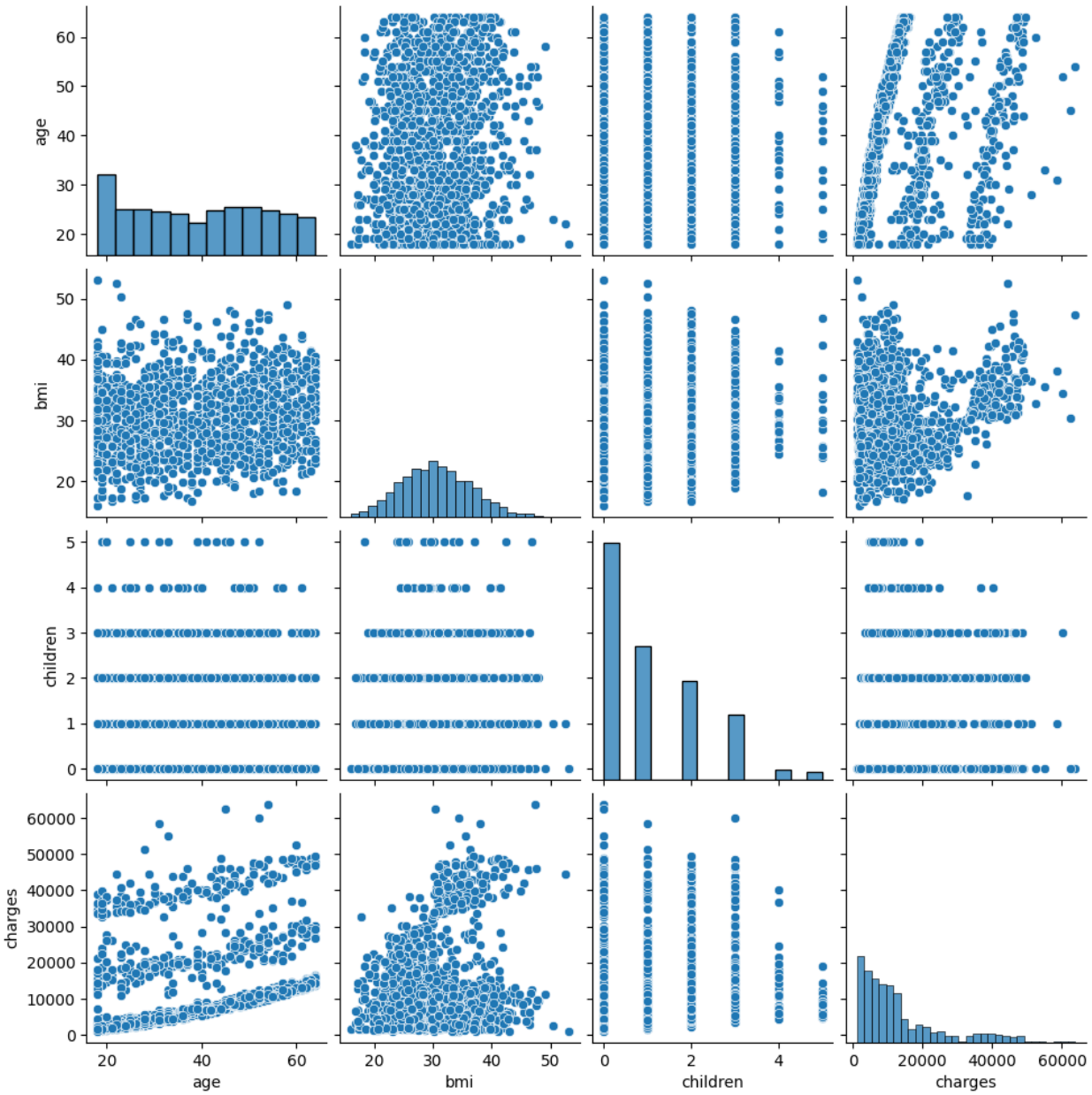


Figure 27: Pair Plot

- Pair plots are a useful tool for visualizing the relationships between multiple variables in a dataset. They help to identify patterns, trends, and associations between variables and are commonly used in exploratory data analysis to understand the structure of the data.

The concept of conditioning variables in a graphics system was pioneered with Trellis graphics,

developed by Rick Becker, Bill Cleveland, and others at Bell Labs. This idea has propagated to various modern graphics systems, such as the lattice and ggplot2 packages in R and the seaborn and Bokeh modules in Python. Conditioning variables are also integral to business intelligence platforms such as Tableau and Spotfire. With the advent of vast computing power, modern visualization platforms have moved well beyond the humble beginnings of exploratory data analysis. However, key concepts and tools developed a half century ago (e.g., simple boxplots) still form a foundation for these systems.

8.6 Key Ideas

- Hexagonal binning and contour plots are useful tools that permit graphical examination of two numeric variables at a time, without being overwhelmed by huge amounts of data.
- Contingency tables are the standard tool for looking at the counts of two categorical variables.
- Boxplots and violin plots allow you to plot a numeric variable against a categorical variable.

9 Summary

Exploratory data analysis (EDA), pioneered by John Tukey, set a foundation for the field of data science. The key idea of EDA is that the first and most important step in any project based on data is to look at the data. By summarizing and visualizing the data, you can gain valuable intuition and understanding of the project. The concepts ranging from simple metrics, such as estimates of location and variability, to rich visual displays that explore the relationships between multiple variables. The diverse set of tools and techniques being developed by the open source community, combined with the expressiveness of the R and Python languages, has created a plethora of ways to explore and analyze data. Exploratory analysis should be a cornerstone of any data science project.